Lecture 2: Approaches to variance estimation

Tim Goedemé
Herman Deleeck Centre for Social Policy

18 January 2018
EUROMOD Winter School, University of Antwerp

Universiteit Antwerpen

Overview

1. Total survey error and the sampling variance
2. The sampling variance
3. The determinants of the sampling variance
4. **Approaches to variance estimation**
5. The ultimate cluster method
6. Analysing subpopulations
7. Comparing point estimates
8. Conclusion

Problem

- We need estimate of sampling variability
- We do not observe:
  - Population distribution
  - Sampling distribution
- How to estimate sampling variance and confidence intervals?

4. Approaches

2 most common approaches:

1. Analytical approaches
2. Replication-based approaches

4. Approaches

- Analytical approaches
  - (non)linear statistics are expressed as totals; linearization (Taylor series expansion);
  - use a standard formula for estimating variance of linearized estimator, asymptotic theory
  - A sampling distribution is assumed to estimate confidence intervals and significance tests

4. Approaches

Inductive approaches:

- Are based on replication of the original sample (or replicate weights)
- random groups method, Balanced repeated replication, Jackknife Repeated Replication, the bootstrap (and replicate weights)

Advantages:

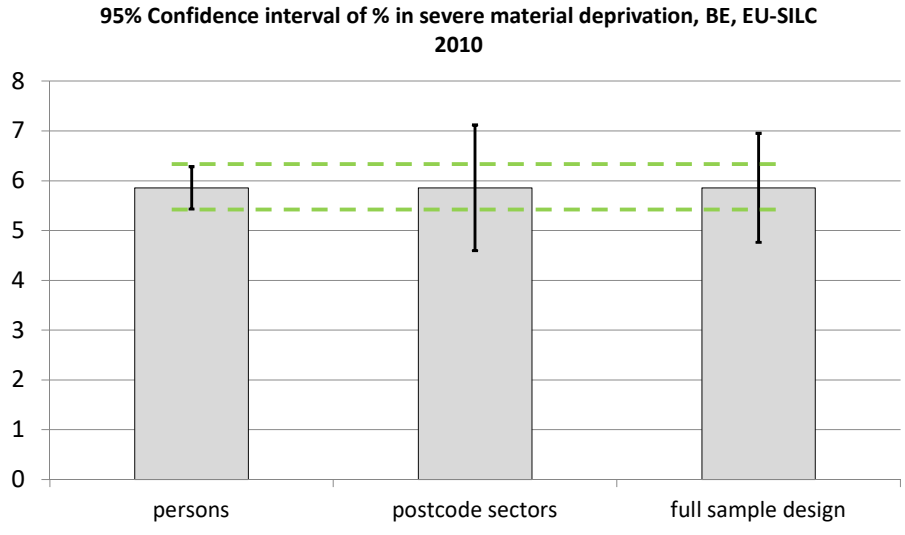
- can be used when no analytical formula (estimation command) is available;
- no assumptions about the shape of the sampling distribution

But computationally intensive; bias

4. Approaches

Whichever approach is chosen, they only work when taking account of the sample design

4. Approaches



5. Ultimate cluster aproach

- How to take account of many different types of sample designs?
- Analytical formula very complex

EU-SILC sample design(s)

Table 26.1: EU-SILC sampling design by country, 2012

Without stratification	
Simple random sampling	Denmark, Iceland, Malta, Norway
Systematic sampling	Sweden
With stratification	
Stratified simple random sampling	Austria, Cyprus, Germany, Lithuania, Luxembourg, Slovakia, Switzerland
Stratified and systematic sampling	Estonia
Stratified two-stage	Croatia, Italy, Latvia, the Netherlands, Portugal, Slovenia
Stratified multi-stage	Belgium, Bulgaria, Czech Republic, Ireland, Greece, Spain, France, Poland, Romania, United Kingdom
Stratified two-phase	Hungary, Finland

Source: Eurostat, 2012 EU-SILC Comparative Quality report (available on CIRCABC).

Source: Berger et al. (2017)

EU-SILC sample design(s)

Special features

- Rotational panel design
  - Sometimes longer panel or pure panel (FR, LU, NO)
  - Rotation at level of PSUs or within PSUs
- Quota sampling in DE until 2008
- (Multiple) changes in sample design over time (esp. HU)
- Calibration on microcensus in NL; on income variables in e.g. SE, FI, …
- Probabilities of selection >1 (e.g. BE)

Universiteit Antwerpen

11

```
. ta country DB060_F, missing
```

DB020	DB060_F -2	1	2	.	Total
AT	12,982	0	0	0	12,982
BE	0	14,346	0	0	14,346
BG	0	12,184	0	0	12,184
CY	12,027	0	0	0	12,027
CZ	0	18,210	0	0	18,210
DE	26,499	0	0	0	26,499
DK	14,078	0	0	0	14,078
EE	15,051	0	0	0	15,051
EL	0	20,995	0	0	20,995
ES	0	0	31,622	0	31,622
FI	0	27,142	0	0	27,142
FR	0	0	26,787	0	26,787
HR	0	14,039	0	0	14,039
HU	0	10,628	12,078	2	22,708
IE	0	14,078	0	0	14,078
IS	8,841	0	0	0	8,841
IT	0	31,605	15,531	0	47,136
LT	11,898	0	0	0	11,898
LU	9,982	0	0	0	9,982
LV	0	14,054	0	0	14,054
MT	11,805	0	0	0	11,805
NL	0	0	24,494	0	24,494
NO	18,419	0	0	0	18,419
PL	0	36,127	0	0	36,127
PT	0	0	17,221	0	17,221
RO	0	17,387	0	0	17,387
SE	14,026	0	0	0	14,026
SI	0	0	27,697	0	27,697
SK	15,711	0	0	0	15,711
UK	0	22,476	0	0	22,476

1 = rotation at PSU level

2 = rotation within PSUs

12

5. Ultimate cluster method

- Only take account of the first stage of the sample design (stratification and clustering)
- Assume there is no subsampling within PSUs: work with observations of PSUs in the condition they are found in the sample
- Assume sampling with replacement

5. Ultimate cluster method

- Why: ease of computation
- Second and subsequent stages add little variance if sampled fraction of PSUs is small (which is often the case)

5. Ultimate cluster method

- Within-stratum variance for the mean (simple random, equal clusters) (Kish, 1965)

$$\mathrm{Var}\,(\bar{y}) = \left(1 - \frac{a}{A}\right)\frac{S_a^2}{a} + \left(1 - \frac{b}{B}\right)\frac{S_b^2}{ab}.$$

- 1-a/A and 1-b/B = FPC

$$s^2(\bar{y}) = \left(1 - \frac{a}{A}\right)\frac{s_a^2}{a} + \left(1 - \frac{b}{B}\right)\frac{a}{A}\frac{s_b^2}{ab}$$

- S²(a)=between-cluster
- S²(b)=within-cluster

5. Ultimate cluster method

Need of good sample design variables to:

- Identify PSUs
- Identify Primary strata
- Take account of calibration (post-stratification, raking)

5. Ultimate cluster method

- Strata with one PSU:
  - One of many PSUs selected (with respondents) -> join similar strata (on sampling frame)
  - Self-representing PSU? -> PSU is stratum, use next stage of sample design as PSUs

5. Ultimate cluster method

Remarks:

- Sample design variables should refer to moment of selection (not interview)
- PSU codes must at least be unique within strata
- Panels: use consistent PSU and strata codes
- Degrees of freedom: #PSUs-#Strata

5. Ultimate cluster method

In Stata

- use sample design variables to identify the sample design
  svyset PSU [pweight = weight], strata(strata)

Subsequently: svy: commands

SPSS: CSPLAN

R: survey package (svydesign and other commands)

SAS: PROC SURVEYFREQ and others

The EU-SILC sample design variables

In EU-SILC, the following sample design variables are available:

- DB050: primary strata (not included in the EU-SILC UDB)
- DB060: primary sampling units
- DB062: secondary sampling units
- DB070: order of selection of primary sampling units

The EU-SILC sample design variables

Especially for earlier waves, quite a few problems.

1/ Missing information

- DB050 lacking
- DB060 lacking (esp. earlier waves, or for 'older' rotational panels)
- With missing DB050: no unique DB060 across strata (e.g. PL, SI); no unique DB070 (UK)
- No 'secondary strata' in case of self-representing PSUs
- When households are split (AT, earlier waves)
- Calibration, imputation

The EU-SILC sample design variables

Especially for earlier waves, quite a few problems.

2/ Moment of selection vs. moment of interview

- DB050 x DB040 (ES, FR, until EU-SILC 2008 at least)
- DB040 as proxy for DB050
- Moving households wrongly received new PSU code (UK)

The EU-SILC sample design variables

Especially for earlier waves, quite a few problems.

3/ Multiple hits => unique DB060 code

- Sampling of PSUs with replacement
- Was not always the case, esp. BE & LV

The EU-SILC sample design variables

Especially for earlier waves, quite a few problems.

4/ Strata with 1 PSU

- Reason was not always clear
  - Self-representing PSUs (e.g. IT, UK, FR)
  - One PSU observed out of many?
- Turn into stratum vs. collapsing strata

The EU-SILC sample design variables

Especially for earlier waves, quite a few problems.

5/ Inconsistent PSU codes

- Across rotational panels
- Across waves

The EU-SILC sample design variables

Example: EU-SILC Belgium. Standard error of difference in AROP60 between two cross-sections with **UDB** (latest releases)

	2008	2009	2010
2009	1.09		
2010	1.10	1.15	
2011	1.26	1.28	1.26

Example: EU-SILC Belgium. Standard error of differer
two cross-sections with **consistent coding of DB060**

	2008	2009	2010
2009	0.62		
2010			
2011	0.87	0.79	0.64

Notes: No difference between the two data sources
sectional estimates! Figures only for illustrative purp

1. Standard error of difference is much smaller with consistent SD variables.
2. Difference with 2011: the longer the time-span, the weaker the covariance (and the larger the standard error) will be

The EU-SILC sample design variables

Especially for earlier waves, quite a few problems.

6/ changes in sample design

- AT: introduced multi-stage with stratification
- NO: abandoned multi-stage design
- HU: change for many rotational panels

- In principle sample elements could have been drawn under different sample designs at the same time

The EU-SILC sample design variables

- Making the best of what we have…
- Do-files at

https://timgoedeme.com/eu-silc-standard-errors/

- Run do-file on D-file of EU-SILC
- Then merge with other EU-SILC files
- svyset psu1 [pw=rb050], strata(strata1)

The EU-SILC sample design variables

Making the best of what we have…

- DB040 as proxy for DB050
  - Regroup if possible
- DB060, or hid (DB030)
- Try to identify / treat special cases
  - Self-representing PSUs (e.g. IT)
  - Make codes consistent / unique across rotation panels
  - Conservative when not both strata and PSUs are available

The EU-SILC sample design variables

Making the best of what we have…

- Each stratum and each PSU unique identifier across entire dataset
- But not consistent across waves…
- Should be made unique across waves (i.e. add year-code) for comparisons between waves.

Steps in analysis

1. definition of the problem
2. Check sample designs and sample design variables (including weights, correlation between weights and variables of interest, ...)
3. Svyset the data and check the sample design, in function of the analysis of interest
4. Inspect missings and imputation -> multiple imputation possible?
5. Inspect outliers and apply proper treatment
6. run proper analysis and interpret results
7. report results, including precision of estimates

Conclusion

Key messages

1. If estimates are based on samples -> estimate and report SEs, CIs & p-values
2. Always take as much as possible account of sample design when estimating SEs, CIs & p-values
3. Never delete observations from the dataset
4. Never simply compare confidence intervals

Literature

- Goedemé, T. (2013) 'How much confidence can we have in EU-SILC?', *Social indicators research*, 110(1): 89-110, doi:10.1007/s11205-011-9918-2
- Heeringa, S. G., West, B. T. and Berglund, P. A. (2010), Applied Survey Data Analysis, Boca Raton: Chapman & Hall/CRC, 467p.
- Wolter, K. M. (2007), Introduction to Variance Estimation, New York: Springer, 447p.
- https://timgoedeme.com/eu-silc-standard-errors/.

Thanks!

tim.goedeme@ua.ac.be